REPRINTED FROM

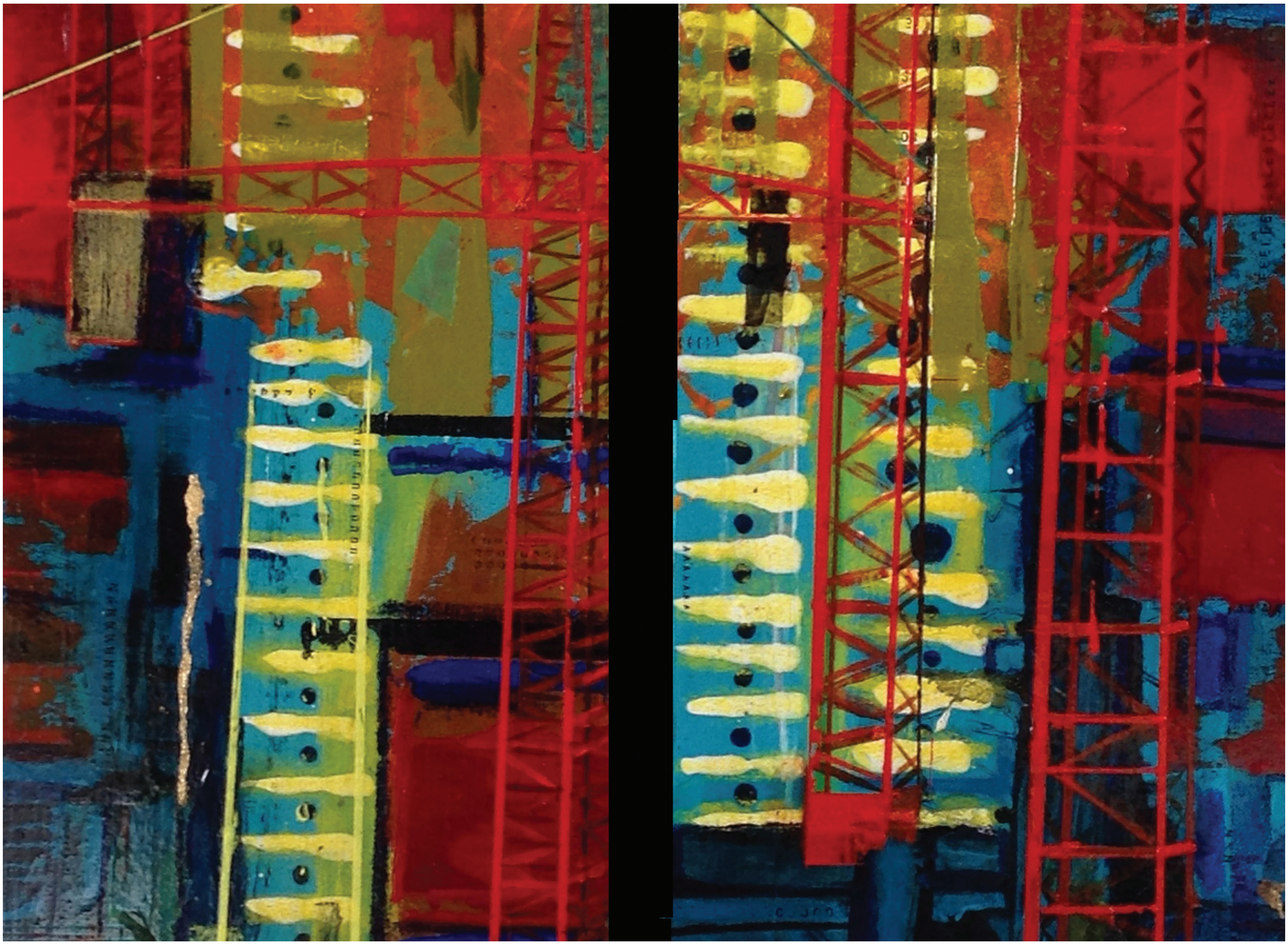
Risk

RISK MANAGEMENT • DERIVATIVES • REGULATION

Risk.net December 2014

Cutting edge
Back-testing expected shortfall

In depth
The shortfalls of expected shortfall



Expected shortfall

The future?

End of the back-test quest?

Ever since regulators suggested replacing value-at-risk with expected shortfall, the industry has been debating how and whether it can be back-tested. Quants at MSCI are proposing three methods. Nazneen Sherif introduces this month's technical articles

From the start, expected shortfall has suffered in comparison with one of the key advantages of the measure it is supposed to be replacing: it cannot be back-tested, critics claimed, while tests of value-at-risk are simple and intuitive.

Regulators have ploughed on regardless. Expected shortfall has been endorsed as VAR's successor in two consultation papers on the *Fundamental review of the trading book* because of its supposed benefits as a measure of tail risk. The widely contested solution to back-testing difficulties is to perform capital calculations using expected shortfall, and then to back-test using VAR. This means the tail is left untested, an outcome regulators concede looks odd (*www.risk.net/2375204*).

There may now be an alternative. In this month's first technical, *Backtesting expected shortfall*, Carlo Acerbi, executive director, and Balázs Székely, senior associate in the analytics research team at MSCI, offer three methods for back-testing expected shortfall, and show they are more efficient than the VAR-based test. They also dismiss arguments based on elicitability – a property expected shortfall was shown to lack in 2011. Elicitability allows a measure to have a scoring function that makes comparison of different models possible, leading many to conclude it would not be back-testable. The authors say that isn't so. They believe elicitability is relevant for model selection and not model testing, and so is not required for regulatory purposes.

Expected shortfall remains difficult to back-test, though, because of the way it is defined – as an average of losses in excess of a given VAR level. In back-testing, the prediction is an entire distribution, but the realisation is a single scenario. Averaging multiple scenarios to calculate expected shortfall means firms hit a wall when trying to back-test.

To date, nearly all attempts at back-testing expected shortfall focused on separately testing VAR and then the size of the exceptions. "Here, one looks at the magnitude of the exceedances beyond VAR, once VAR is already back-tested and you know it is correct. Obviously, it's tough to take a decision when one of the two parts fails," says Acerbi.

Of the three tests the authors propose, the second – which they believe to be the easiest to implement – jointly tests for frequency and magnitude of the VAR exceptions. They exploit the lesser-known unconditional expectation version of the expected shortfall formula, which eliminates the need to check VAR exceptions beforehand.

The result is a non-parametric test, free from assumptions on distribution, with greater ability to detect an effect than the VAR test. It also stores two values – magnitude and frequency-per-day – eliminating the need for Monte Carlo simulations for most practical cases because the thresholds separating the categories of the current regulatory traffic-light system for back-testing VAR remain more or less fixed.

"Expected shortfall has better properties than VAR, so if it's back-testable, there is no reason to use the VAR test anymore, which was a debatable choice in the first place," says Fabrizio Anfuso, head of counterparty credit risk back-testing methodology at Credit Suisse's investment bank.

It's too soon to break out the champagne, however. The trading book review also attempts to capture liquidity risks by introducing a spread of different time horizons for individual risk factors, which would sink any attempt to back-test, Acerbi warns: "Back-testing any measure, including VAR, on asynchronous time horizons spoils a fundamental assumption in back-testing – time independence of different observations. No back-testing method, be it for VAR or expected shortfall, fits within that framework to the best of my knowledge."

"Expected shortfall has better properties than VAR, so if it's back-testable, there is no reason to use the VAR test anymore, which was a debatable choice in the first place"
Fabrizio Anfuso, Credit Suisse

Eduardo Epperlein, global head of risk methodology at Nomura, echoes the point: "If we have different horizons for different risk factors, then there is the question of what the correlation between them is, and this is not something we can easily back-test."

The inclusion of stressed scenarios in capital calculations poses another hurdle. "The basic back-test addresses the problem of how well you predict tomorrow's one-day expected shortfall, but it does not say whether you did a better or worse job at estimating a stressed window and its respective stressed VAR," says Epperlein. University of Toronto's John Hull and Alan White raised similar issues in an opinion piece last month (*Risk* November 2014, *www.risk.net/2375185*).

Some see expectiles as a good alternative to expected shortfall because they are both elicitable and back-testable. However, their dependence on both profits and losses means two banks claiming different profits could have different capital charges even if they have the same loss history – something regulators would probably not be thrilled about.

"In my opinion, expectiles aren't any better than expected shortfall at managing risk. I don't think they are a particularly useful avenue," says Richard Martin, principal at Apollo Global Management.

In our second technical, *KVA: Capital valuation adjustment by replication*, Andrew Green, head of the credit valuation adjustment/funding valuation adjustment quantitative research team at Lloyds Banking Group and Chris Kenyon, a director in the same team, introduce a new valuation adjustment based on regulatory capital that has been gaining traction recently among quants, and propose an efficient way of calculating it. **R**

Back-testing expected shortfall

The discovery that expected shortfall (ES) is not elicitable propagated the belief that it could not be back-tested and aroused a number of criticisms of the Basel Committee's adoption of ES over value-at-risk. In this article, Carlo Acerbi and Balázs Székely propose three back-testing methodologies for ES that are more powerful than the Basel VAR test, and observe that elicitability is irrelevant when it comes to the choice of a regulatory risk standard

Risk professionals had never heard of elicitability before 2011, when Gneiting (2011) proved that expected shortfall (ES) is not elicitable, unlike value-at-risk. This result sparked a confusing debate.

Put simply, a statistic $\psi(Y)$ of a random variable Y is said to be elicitable if it minimises the expected value of a scoring function S:

$$\psi = \arg\min_x \mathbb{E}[S(x, Y)]$$

Given a history of point predictions x_t for the statistics and realisations y_t of the random variable, this provides a natural way to evaluate the forecast model, by requiring the mean score:

$$\bar{S} = \frac{1}{T}\sum_{t=1}^{T} S(x_t, y_t)$$

to be as low as possible. The mean and the median represent popular examples, minimising the mean square and absolute error, respectively. The α-quantile, hence VAR, is also elicitable, with score function $S(x, y) = ((x > y) - \alpha)(x - y)$, a well-known fact in quantile estimation.

The discovery that ES cannot be elicited led many to conclude that it would not be back-testable (see, for example, Carver 2013) and sounded like the formal proof of a fact that had long been suspected. It is true that the absence of a convincing back-test has long been the last obstacle for ES on its way to Basel.

In October 2013, a consultation paper from the Basel Committee on Banking Supervision (2013) opted to replace VAR with ES for determining the capital charge of internal models, but VAR was kept as the measure for use in back-testing in the usual way. The change was criticised based on the alleged impossibility of back-testing ES, which was interpreted as a sign that there is something inherently wrong with this risk measure.

Not everyone, however, was convinced. If elicitable means back-testable, where does that leave the few (but valuable) works on ES back-testing, such as Kerkhof & Melenberg (2004), that make conclusions like the following: 'contrary to common belief, ES is not harder to backtest than VAR.…Furthermore, the power of the test for ES is considerably higher'? And what should we do with variance, given that it is not elicitable either? And why has VAR never been back-tested by exploiting its elicitability? At a certain point, some dissenting voices started to emerge (Emmer, Kratz & Tasche 2013; Tasche 2013).

In what sense, if any, is it more difficult to back-test ES than VAR? For fundamental reasons? Because of practical aspects? Is it the power of the test? Model risk? To address these questions we introduce some statistical tests for ES and compare them with VAR back-tests. We restrict our choice to tests that are non-parametric and free from distributional assumptions other than continuity, which is a necessary condition for any application in banking regulation.

An extended version of this paper with all proofs and experiments is Acerbi & Székely (2014).

Back-testing ES

We adopt a standard hypothesis testing framework for unconditional coverage of ES analogous to the standard Basel VAR setting. We assume that independence of arrival of tail events is tested separately, typically just by visual inspection of VAR exception clusters. This is still the preferred practice in the industry as it provides better insight than proposed tests such as those in Christoffersen (1998).

We assume that every day $t = 1, \dots, T$, X_t represents a bank's profit and loss distributed along a real (unknowable) distribution F_t and it is forecasted by a model predictive distribution P_t conditional on previous information used to compute $\mathrm{VAR}_{\beta,t}$ and $\mathrm{ES}_{\alpha,t}$ as defined by (see Acerbi & Tasche 2002):

$$\mathrm{ES}_{\alpha,t} = -\frac{1}{\alpha}\int_0^{\alpha} P_t^{-1}(q)\,\mathrm{d}q \tag{1}$$

The random variables $\vec{X} = \{X_t\}$ are assumed to be independent but not identically distributed. We do not restrict the variability of F_t and P_t over time in any respect. We will denote by $\mathrm{VAR}^F_{\alpha,t}$ and $\mathrm{ES}^F_{\alpha,t}$ the value of the risk measures when $X \sim F$.

We assume in what follows that the distributions are continuous and strictly increasing, in which case ES can be expressed as:

$$\mathrm{ES}_{\alpha,t} = -\mathbb{E}[X_t \mid X_t + \mathrm{VAR}_{\alpha,t} < 0] \tag{2}$$

and VAR is uniquely defined as $\mathrm{VAR}_{\beta,t} = -P_t^{-1}(\beta)$. In real cases, this assumption is completely innocuous.[1]

Without loss of generality, in our numerical examples we will use $T = 250$, $\beta = 1\%$ and $\alpha = 2.5\%$, which is the relevant case from Basel Committee on Banking Supervision (2013). $\mathrm{ES}_{2.5\%}$ was correctly chosen by the Basel Committee to equal $\mathrm{VAR}_{1\%}$ for Gaussian tails, and to penalise heavier tails. This is analogous to replacing '50' with '80' on road signs when switching from miles per hour to kilometres per hour.

Our null hypothesis generically assumes the prediction is correct, while the alternative hypotheses are chosen to be only in the direction of risk underestimation. This is again in line with the Basel VAR test, which is meant to detect only excesses of VAR exceptions. We formulate more precise test-specific versions of H_0 and H_1 below. Concrete H_1 examples will be analysed in a later section to compute the power

[1] We also assume that $\mathrm{VAR}_\alpha > 0$ as it happens in a realistic portfolio profit and loss distribution.

of tests in selected cases, similar to the approach followed in Basel Committee on Banking Supervision (2013, table 1) for different levels of VAR coverage mismatch.

Test 1: testing ES after VAR Our first test is inspired by the conditional expectation (2), from which we can easily derive:

$$\mathbb{E}\left[\frac{X_t}{\mathrm{ES}_{\alpha,t}} + 1 \,\middle|\, X_t + \mathrm{VAR}_{\alpha,t} < 0\right] = 0 \tag{3}$$

If $\mathrm{VAR}_{\alpha,t}$ has been tested already, we can separately test the magnitude of the realised exceptions against the model predictions. Defining $I_t = (X_t + \mathrm{VAR}_{\alpha,t} < 0)$, the indicator function of an α-exception, we define the test statistics:

$$Z_1(\vec{X}) = \frac{\sum_{t=1}^{T}(X_t I_t/\mathrm{ES}_{\alpha,t})}{N_T} + 1 \tag{4}$$

if $N_T = \sum_{t=1}^{T} I_t > 0$.

For this test, we choose a null hypothesis:

$$\mathrm{H}_0\colon\ P_t^{[\alpha]} = F_t^{[\alpha]},\ \forall t$$

where $P_t^{[\alpha]}(x) = \min(1, P_t(x)/\alpha)$ is the distribution tail for $x < -\mathrm{VAR}_{\alpha,t}$. The alternatives are:

$$\begin{aligned}\mathrm{H}_1\colon\ &\mathrm{ES}_{\alpha,t}^F \geqslant \mathrm{ES}_{\alpha,t},\ \text{for all } t, \text{ and} > \text{for some } t\\ &\mathrm{VAR}_{\alpha,t}^F = \mathrm{VAR}_{\alpha,t},\ \text{for all } t\end{aligned}$$

We see that the predicted VAR_α is still correct under H_1, in line with the idea that this test is subordinated to a preliminary VAR test. This test is in fact completely insensitive to an excessive number of exceptions as it is an average taken over the exceptions themselves.

Under these conditions, $\mathbb{E}_{\mathrm{H}_0}[Z_1 \mid N_T > 0] = 0$ and $\mathbb{E}_{\mathrm{H}_1}[Z_1 \mid N_T > 0] < 0$. So, the realised value $Z_1(\vec{x})$ is expected to be zero, and it signals a problem when it is negative.

Dividing (3) by $\mathrm{ES}_{\alpha,t}$ was unnecessary. Normalising by another statistic of P_t, or not normalising at all, would have given other legitimate tests. Our choice was made to obtain a dimensionless test statistic and to control for heteroscedasticity.

Variations of this test have already appeared in the literature several times. For instance, Mcneil & Frey (2000) proposed something similar in a Garch-EVT context.

Test 2: testing ES directly A second test follows from the unconditional expectation:

$$\mathrm{ES}_{\alpha,t} = -\mathbb{E}\left[\frac{X_t I_t}{\alpha}\right] \tag{5}$$

which suggests the following definition:

$$Z_2(\vec{X}) = \sum_{t=1}^{T}\frac{X_t I_t}{T\alpha\,\mathrm{ES}_{\alpha,t}} + 1 \tag{6}$$

Appropriate hypotheses for this test are:

$$\begin{aligned}\mathrm{H}_0\colon\ &P_t^{[\alpha]} = F_t^{[\alpha]},\ \forall t\\ \mathrm{H}_1\colon\ &\mathrm{ES}_{\alpha,t}^F \geqslant \mathrm{ES}_{\alpha,t},\ \text{for all } t \text{ and} > \text{for some } t\\ &\mathrm{VAR}_{\alpha,t}^F \geqslant \mathrm{VAR}_{\alpha,t},\ \text{for all } t\end{aligned}$$

We again have $\mathbb{E}_{\mathrm{H}_0}[Z_2] = 0$ and $\mathbb{E}_{\mathrm{H}_1}[Z_2] < 0$. Remarkably, these results do not require independence of the X_ts. Furthermore, the test can be immediately extended to general, non-continuous distributions by replacing I_t with:

$$\begin{aligned}I_t' &= (X_t + \mathrm{VAR}_{\alpha,t} < 0)\\ &\quad + \frac{\alpha - \mathrm{Prob}[X_t + \mathrm{VAR}_{\alpha,t} < 0]}{\mathrm{Prob}[X_t + \mathrm{VAR}_{\alpha,t} = 0]}(X_t + \mathrm{VAR}_{\alpha,t} = 0)\end{aligned}$$

See Acerbi & Tasche (2002, equation (4.12)).

Test 2 jointly evaluates the frequency and magnitude of α-tail events as shown by the relationship:

$$Z_2 = 1 - (1 - Z_1)\frac{N_T}{T\alpha} \tag{7}$$

remembering that $\mathbb{E}_{\mathrm{H}_0}[N_T] = T\alpha$.

We remark that both test 1 and test 2 might have been defined under the weaker null hypothesis:

$$\begin{aligned}\mathrm{H}_0'\colon\ &\mathrm{ES}_{\alpha,t}^F = \mathrm{ES}_{\alpha,t},\ \text{for all } t\\ &\mathrm{VAR}_{\alpha,t}^F = \mathrm{VAR}_{\alpha,t},\ \text{for all } t\end{aligned} \tag{8}$$

all the above results holding true. This choice would have not been sufficient, however, to simulate the test statistics and compute p-values (see the section on test significance and power).

Test 3: estimating ES from realised ranks Following Berkowitz (2001) it is possible to back-test the tails of a model by checking if the observed ranks $U_t = P_t(X_t)$ are independent and identically distributed (iid) $U(0,1)$, as they should be if the model distribution is correct. To convert this idea into a specific test for ES, we must assign to each quantile its dollar importance, which depends on the shape of the tail itself. To this end, denoting by:[2]

$$\widehat{\mathrm{ES}}_\alpha^{(N)}(\vec{Y}) = -\frac{1}{[N\alpha]}\sum_{i}^{[N\alpha]} Y_{i:N} \tag{9}$$

an ES estimator based on a vector of N iid draws $\vec{Y} = \{Y_i\}$, we define:

$$Z_3(\vec{X}) = -\frac{1}{T}\sum_{t=1}^{T}\frac{\widehat{\mathrm{ES}}_\alpha^{(T)}(P_t^{-1}(\vec{U}))}{\mathbb{E}_V[\widehat{\mathrm{ES}}_\alpha^{(T)}(P_t^{-1}(\vec{V}))]} + 1 \tag{10}$$

where $\vec{V}$ are iid $U(0,1)$. The idea is that the entire vector of ranks $\vec{U} = \{U_t\}$ is reused to estimate ES for every past day t, and the result is then averaged over the entire period.[3] In the denominator we have not $\mathrm{ES}_{\alpha,t}$ but a finite sample estimate to compensate for the bias of estimator (9). The denominator can be computed analytically via:

$$\begin{aligned}&\mathbb{E}_V[\widehat{\mathrm{ES}}_\alpha^{(T)}(P_t^{-1}(\vec{V}))]\\ &\quad = -\frac{T}{[T\alpha]}\int_0^1 I_{1-p}(T - [T\alpha], [T\alpha])P_t^{-1}(p)\,\mathrm{d}p\end{aligned} \tag{11}$$

where the function $I_x(a,b)$ is a regularised incomplete beta function.

[2] $[x]$ is the integer part of x and $Y_{i:N}$ denotes order statistics.
[3] We could just as well have chosen the distribution $P_{t^\star}$ of a specific day: the last one, $t^\star = T$, for example.

We have $\mathbb{E}_{H_0}[Z_3] = 0$ and $\mathbb{E}_{H_1}[Z_3] < 0$ also in this case. However, the hypotheses this time involve the entire distributions:

$$\begin{aligned} H_0 &: \ P_t = F_t, \ \forall t \\ H_1 &: \ P_t \succeq F_t, \ \text{for all } t \text{ and} \succ \text{for some } t \end{aligned}$$

where $(\succeq) \succ$ denotes (weak) first-order stochastic dominance.

Test 3 is less natural than tests 1 and 2, but it is very general. A similar test may be designed for any other conceivable statistics for which an estimator is available.

Test significance and power

Significance For all tests $Z = Z_i$ we simulate the distribution P_Z under H_0 to compute the p-value $p = P_Z(Z(\vec{x}))$ of a realisation $Z(\vec{x})$:

$$\begin{aligned} &\text{simulate independent } X_t^i \sim P_t, \ \forall t, \ i = 1, \dots, M \\ &\text{compute } Z^i = Z(\vec{X}^i) \\ &\text{estimate } p = \sum_{i=1}^{M} (Z^i < Z(\vec{x}))/M \end{aligned} \tag{12}$$

where M is a suitably large number of scenarios. Given a preassigned significance level ϕ, the test is finally accepted or rejected if $p \gtrless \phi$.

From the above procedure, we see that while it is sufficient to record a single number I_t per day to back-test VAR exceptions, it may be necessary to retain a memory of all predictive distributions P_t to back-test ES.

In reality, for Z_1 and Z_2 it is sufficient to record only the α-tail $P_t^{[\alpha]}$ of the predictive distributions, because $X_t I_t$ can be simulated after $I_t \sim \text{Bernoulli}(\alpha)$. We will see in a later section that in fact Z_2 lends itself to implementations that do not even require the the predictive distributions to be recorded.

Storage of more information (a cumulative distribution function per day) is the only difference between back-testing ES and VAR; this is only a practical difference and it poses no technological challenge.

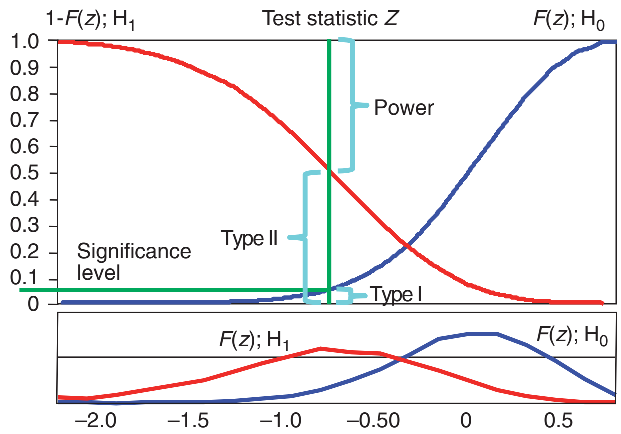
Power In the following subsections, we run a number of experiments to evaluate the power of the $\text{ES}_{2.5\%}$ tests and compare it with the power of the Basel $\text{VAR}_{1\%}$ test under selected hypotheses. The examples are based on Student t distributions, which allow us to span all possible fat-tails indexes. Figure 1 shows how to read the results of every experiment. The green vertical lines in the plots correspond exactly to the 5% and 10% significance levels, while the black vertical lines are the corresponding closest discrete levels attainable by the VAR test.

The results are summarised in tables in which the left part describes the setup of H_0 and H_1 and the right part the power of the tests. Every row in the tables corresponds to one of the significance levels attainable by the VAR test.

Z_1 is not applicable to the examples in which $\text{VAR}_{2.5\%}$ varies across the alternatives.

Scaled distributions: ES coverage We assume that H_1 is a rescaled version of the H_0 distribution: $F(x) = P(x/\gamma)$, $\gamma > 1$.

1 Power and significance for a test under specific H_0s and H_1s



The top plot shows $F(z)$ under H_0 and $1 - F(z)$ under H_1 for the test variable Z. All vertical lines intercept the two curves at type I probability (the significance level) and type II probability, and display the test power. The bottom plot shows the densities under H_0 and H_1; more powerful tests result in more distinct curves

A. Power of multiple tests for scaled distributions with different ES coverage, as explained in the section ‘Scaled distributions: ES coverage’

ν	α' (%)	$\text{VAR}_{1\%}$	$\text{ES}_{2.5\%}$
5	2.5	3.36	3.52
	5	4.10	4.29
	10	5.14	5.38
100	2.5	2.36	2.37
	5	2.68	2.70
	10	3.16	3.18

				Power		
ν	Significance level (%)	Coverage α' (%)	Scale γ (%)	Z_2 (%)	Z_3 (%)	$\text{VAR}_{1\%}$ (%)
5	4.1	5	21.9	51.8	25.2	37.4
		10	53	98.5	78.3	93.5
	10.6	5	21.9	69.0	46.4	55.7
		10	53	99.5	92.9	97.3
100	4.0	5	13.7	47.1	39.0	38.8
		10	34	97.4	94.1	94.2
	10.8	5	13.7	64.7	59.1	56.3
		10	34	99.0	98.1	97.6

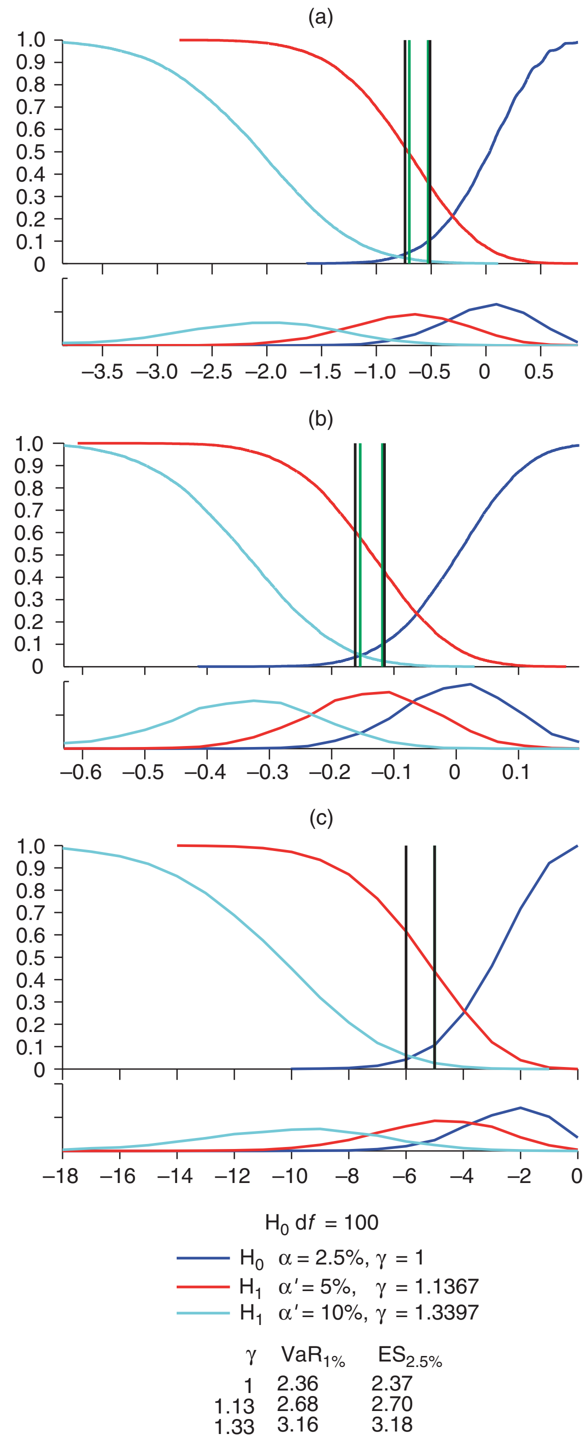
We assume certain levels of ES coverage mismatch, assuming $\text{ES}_\alpha^P = \text{ES}_{\alpha'}^F$ for $\alpha' = 5\%, 10\%$, so that $\gamma = \text{ES}_\alpha^P / \text{ES}_{\alpha'}^P$. The results are shown in table A and figure 2, in which H_0 is chosen to be $\nu = 100$ and $\nu = 5$, respectively.

In these cases, Z_2 clearly outperforms the VAR test in terms of power. Z_3, though, has slightly less power for the smaller tail index.

Student t distributions We choose H_1 to be a Student t distribution with smaller ν than H_0. Note that in this way the variance will also be larger, as $\sigma^2 = \nu/(\nu - 2)$. We analyse two H_0 tail indexes $\nu = 10$, $\nu = 100$ in table B. If $\nu = 100$ (respectively $\nu = 10$), then power is computed for H_1 corresponding to $\nu = 10, 3$ (respectively $\nu = 5, 3$). The results are summarised in the top half of table B.

In this case, Z_3 is the most powerful, followed by Z_2 and by the VAR test.

2 Different ES coverage: (a) $ES_{2.5\%}$, test 2; (b) $ES_{2.5\%}$, test 3; (c) $VAR_{1\%}$, number of exceedances



H_0: Student t with $\nu = 100$

■ **Normalised Student t distributions** We repeat the previous experiment using normalised Student t distributions with unit variance. In this case, the difference between H_0 and H_1 is only due to tail properties and not larger variance. The results are reported in the bottom half of table B.

B. Power of multiple tests for varying tail indexes in H_1: Student t and normalised Student t distributions are investigated

	ν	$VAR_{1\%}$	$ES_{2.5\%}$
Student t	100	2.36	2.37
	10	2.76	2.81
	5	3.36	3.52
	3	4.54	5.04
Normalised Student t	100	2.34	2.35
	10	2.47	2.52
	5	2.60	2.72
	3	2.62	2.90

	ν in H_0	Significance level (%)	ν in H_1	Power Z_2 (%)	Power Z_3 (%)	Power $VAR_{1\%}$ (%)
Student t	10	4.0	5	43.4	48.9	37.7
			3	92.3	94.0	87.1
		10.6	5	61.3	66.1	55.5
			3	96.5	97.1	93.5
	100	4.1	10	40.9	54.8	38.2
			3	99.3	99.8	98.5
		10.4	10	57.7	67.7	56.3
			3	99.6	99.9	99.5
Normalised Student t	10	4.4	5	7.8	18.7	9.0
			3	8.6	31.4	7.4
		11.2	5	16.5	30.6	18.7
			3	16.0	41.1	16.8
	100	4.4	10	8.2	22.1	10.5
			3	12.3	49.1	12.0
		11.0	10	17.9	34.3	21.6
			3	20.5	56.6	24.5

This case is particularly subtle. Both Z_2 and the VAR test display very little power at all, with VAR doing slightly better. Z_3, by contrast, performs quite well.

■ **Fixed $VAR_{2.5\%}$ Student t distributions** In order to analyse Z_1, we repeat the experiments using Student t and normalised Student t, and also shifting the H_1 distributions in such a way as to leave $VAR_{2.5\%}$ unchanged. The results are reported in table C.

Z_2 and the VAR test also display modest power in this case. On the other hand, both Z_1 and Z_3 perform very well.

■ **A comment on the results** In these experiments, and others that have been performed, Z_2 has proved to be the most powerful in the case of alternative hypotheses with different volatility, while Z_3 and Z_1 were the most powerful in the case of a different tail index. The VAR exceptions test is generally significantly less powerful.

■ **Avoiding storage of predictive distributions for Z_2** The critical levels for Z_2 display remarkable stability across different distribution types. Table D illustrates the levels for 5% and 0.01% significance (the levels used in the Basel traffic light mechanism) for Student t distributions with different νs and means.

It is clear that a traffic light mechanism based on Z_2 with fixed levels $Z_2^\star = -0.7$ and $Z_2^\star = -1.8$ would do perfectly in all occasions. Note that the ± 1 location shifts span an unrealistically large region for a real profit and loss distribution, which is expected to be centred around zero. Note also that the thresholds deviate significantly only for dramatically heavy-tailed distributions, with $\nu = 3$, and that in this case the proposed test would be more penalising, which is probably a good thing, given that such tails represent a problem in and of themselves.

C. Power of multiple tests in the experiment of the section titled 'Fixed $\text{VAR}_{2.5\%}$ Student t distributions'; similar to table B but with distributions with fixed $\text{VAR}_{2.5\%}$

	ν	$\text{VAR}_{1\%}$	$\text{ES}_{2.5\%}$
Student t	100	2.36	2.37
	10	2.51	2.57
	5	2.77	2.93
	3	3.34	3.84
Normalised Student t	100	2.34	2.35
	10	2.44	2.49
	5	2.57	2.70
	3	2.74	3.03

	ν in H_0	Significance level (%)	ν in H_1	Power Z_1 (%)	Z_2 (%)	Z_3 (%)	$\text{VAR}_{1\%}$ (%)
Student t	10	4.1	5	28.6	11.1	27.4	12.0
			3	72.7	28.8	62.8	24.9
		10.7	5	43.7	20.4	39.1	24.4
			3	82.2	39.8	70.6	41.6
	100	4.3	10	28.2	7.7	25.1	11.0
			3	91.7	38.5	79.5	33.6
		10.9	10	43.2	15.9	36.3	22.1
			3	94.4	49.1	83.3	50.8
Normalised Student t	10	4.2	5	20.1	7.9	19.0	8.7
			3	44.7	16.0	39.3	13.8
		11.4	5	33.5	16.8	29.9	18.8
			3	58.5	27.5	50.2	26.9
	100	4.1	10	21.2	6.0	18.9	8.3
			3	70.3	19.6	59.8	20.7
		11.1	10	35.2	13.7	29.4	18.6
			3	79.2	31.4	67.4	35.9

The important fact behind this stability is that for implementing Z_2 there is effectively no need to do a Monte Carlo test and therefore no need to store predictive distributions. Testing Z_2 requires recording only two numbers per day: the magnitude $X_t I_t$ of a $\text{VAR}_{\alpha,t}$ exception, and the predicted $\text{ES}_{\alpha,t}$.

Back to elicitability

Now that we have seen that elicitability is not necessary for back-testing, we argue something else: that in fact, elicitability has nothing to do with back-testing.

Model selection, not model testing Elicitability allows us to compare in a natural way (yet not the only possible way) different models that forecast statistics in the exact same sequence of events, while recording only point predictions. For instance, if a bank has multiple VAR models in place for its profit and loss, the mean score can be used to select the best in class. But this is model selection, not model testing. It is a relative ranking not an absolute validation.

Regulators, by contrast, need to validate individual models from different banks on an absolute scale. For this purpose elicitability is of no use. A hypothesis test based on elicitability would still require either collection of the predictive distributions or strong distributional assumptions, with no guarantee of better power *a priori*.

It is not a coincidence then that, despite VAR being elicitable, VAR back-tests are still based on counting exceptions. If these tests are simple and entail the recording of just one number, it is not because

D. 5% and 0.01% significance thresholds for Z_2 across Student t distributions with different νs and locations

	5% significance			0.01% significance		
	Location			Location		
ν	−1	0	1	−1	0	1
3	−0.78	−0.82	−0.88	−3.9	−4.4	−5.5
5	−0.72	−0.74	−0.78	−1.9	−2.0	−2.3
10	−0.70	−0.71	−0.74	−1.8	−1.9	−1.9
100	−0.70	−0.70	−0.72	−1.8	−1.8	−1.9
Gaussian	−0.70	−0.70	−0.72	−1.8	−1.8	−1.9

VAR is elicitable, but because quantiles define a Bernoulli random variable. No other elicitable statistic does this.

Expectiles Expectiles have recently attracted a lot of interest (see, for example, Martin 2014) because they are the only coherent law-invariant measure of risk that is also elicitable (Bellini *et al* 2014; Ziegel 2014). But while, as we have seen, the absence of elicitability is not a serious problem for a regulatory risk standard, the absence of comonotonic additivity certainly is. An expectile ρ will tell you that a long position in a call option C is partially hedged by a long (yes, long) position in the underlying stock S: $\rho(C+S) < \rho(C) + \rho(S)$.

The class of comonotonic additive coherent measures of risk of law-invariant type has been completely classified and it coincides with spectral measures of risk (see Acerbi 2007), which contain ES as the most popular example. Alternative choices that are not law invariant belong to the realm of measures based on stress tests – which, incidentally, is an avenue the Fed seems to be considering with increasing interest.

Joint elicitability of ES and VAR An intuitive, if not rigorous, way to understand why ES is not elicitable is to notice that there exists no expression of the type:

$$\mathbb{E}[L(X, \text{ES})] = 0$$

where L is a function involving only a random variable and its ES. If such a function existed we could interpret it as:

$$L(X, \text{ES}) = \left.\frac{\partial S(X,e)}{\partial e}\right|_{e=\text{ES}}$$

and integrate it with respect to e to build a scoring function S that elicits ES. However, there exist null expectations that involve both ES and VAR. For instance:

$$\begin{aligned} \mathbb{E}[(X+\text{ES})(X+\text{VAR}<0)] &= 0 \\ \mathbb{E}[X(X+\text{VAR}<0)+\alpha\text{ES}] &= 0 \end{aligned} \tag{13}$$

It is therefore clear that if there is a chance to build a scoring function for ES, this needs to involve VAR as well. Starting from the above expressions, it is in fact not difficult to construct a one-parameter family of scoring functions:

$$\begin{aligned} S^W(v,e,x) = {} & \alpha e^2/2 + W\alpha v^2/2 - \alpha e v \\ & + (e(v+x) + W(x^2 - v^2)/2)(x+v<0) \end{aligned} \tag{14}$$

for every $W \in \mathbb{R}$, that jointly elicit VAR and ES:

$$\{\text{VAR}, \text{ES}\} = \underset{v,e}{\arg\min}\, \mathbb{E}[S^W(v,e,X)] \tag{15}$$

under the condition that $\mathrm{VAR} \cdot W > \mathrm{ES}$. Note that, for any fixed W we can imagine a bizarre distribution (eg, a $\nu = 1 + \epsilon$ Student t with $\epsilon > 0$ small enough) that violates this condition, so, strictly speaking, this is not a mathematical proof of 2-elicitability in the sense of Lambert, Pennock & Shoham (2008), as was given for variance and mean. However, from a practical point of view, it is easy to choose a value for W that is large enough for any specific case at hand.

As an aside, we observe that theoretical results showing that a measure is not elicitable may still not preclude it being elicitable in practice. We still do not know whether VAR and ES are jointly elicitable, and we would not be surprised to discover that they are not, but we already know that in practice they are.

We note that this result opens up new ways of setting up selections for ES models, but in light of the observations made in the section on model selection, it does not add anything to ES as a candidate for regulatory standards.

Conclusions

Expected shortfall can be back-tested. The most important contribution of our work is to define three ES back-test methods that are non-parametric, distribution-independent and do not assume any asymptotic convergence. The tests are easy to implement and generally display better power than the standard Basel VAR back-test. The only additional complexity they bring about is the necessity to record the predicted cumulative distribution function day by day, and even this is unnecessary for Z_2, which exhibits remarkable stability in the critical levels across different tail shapes.

The elicitability of a risk measure is not relevant for absolute model validation. This property is useful for relative comparison of different models forecasting the same process, namely for model selection. The non-elicitability of a risk measure does not preclude the possibility of back-testing it efficiently, and the elicitability of VAR will never provide a better alternative to back-testing it just by counting exceptions.

We provide some insight into why ES is not individually elicitable. From this fact, we learn how to build a scoring functional that jointly elicits ES and VAR. The result is new and generally important for ES model selection, but we do not think it will affect, in any respect, the regulatory debate around VAR and ES.

We believe that Z_1 in tandem with the standard Basel VAR back-test or, alternatively, Z_2 alone represent valid proposals for back-testing models for ES-based regulation. Z_3 is also a valid test, but it seems to be more appropriate as a complementary test solely for detecting tail index misspecification. **R**

Carlo Acerbi is an executive director in the analytics research team at MSCI Inc. Balázs Székely is a senior associate in the same team. **The authors are indebted to Tilmann Gneiting, Dirk Tasche and an anonymous referee for many corrections and enriching discussions. We are grateful to Imre Kondor for spurring this research and for organising the 'International Workshop on Systemic Risk and Regulatory Market Risk Measures' in Pullach, where a preliminary version of this work was first presented.** **Email: carlo.acerbi@msci.com, balazs.szekely@msci.com**

REFERENCES

Acerbi C, 2007
Coherent measures of risk in everyday market practice
Quantitative Finance 7(4), pages 359–364

Acerbi C and B Székely, 2014
Backtesting expected shortfall (extended version)
MSCI Research Paper

Acerbi C and D Tasche, 2002
On the coherence of expected shortfall
Journal of Banking & Finance 26, pages 1487–1503

Basel Committee on Banking Supervision, 2013
Fundamental review of the trading book: a revised market risk framework
Second Consultative Paper

Bellini F, B Klar, A Müller and E Rosazza Gianin, 2014
Generalized quantiles as risk measures
Insurance: Mathematics and Economics 54, pages 41–48

Berkowitz J, 2001
Testing density forecasts, applications to risk management
Journal of Business & Economic Statistics 19, pages 465–474

Carver L, 2013
Mooted VAR substitute cannot be back-tested, says top quant
www.risk.net/2253463

Christoffersen P, 1998
Evaluating interval forecasts
International Econonomic Review 39, pages 841–862

Emmer S, M Kratz and D Tasche, 2013
What is the best risk measure in practice? A comparison of standard measures
Working paper

Gneiting T, 2011
Making and evaluating point forecasts
Journal of the American Statistical Association 106, pages 746–762

Kerkhof J and B Melenberg, 2004
Backtesting for risk-based regulatory capital
Journal of Banking & Finance 28, pages 1845–1865

Lambert N, DM Pennock and Y Shoham, 2008
Eliciting properties of probability distributions
In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC '08)*, pages 129–138

Martin R, 2014
Expectiles behave as expected
Risk June, pages 79–83

McNeil AJ and R Frey, 2000
Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach
Journal of Empirical Finance 7, pages 217–300

Tasche D, 2013
Expected shortfall is not elicitable. So what?
Presentation at Imperial College, London

Ziegel JF, 2014
Coherence and elicitability
Mathematical Finance, to appear

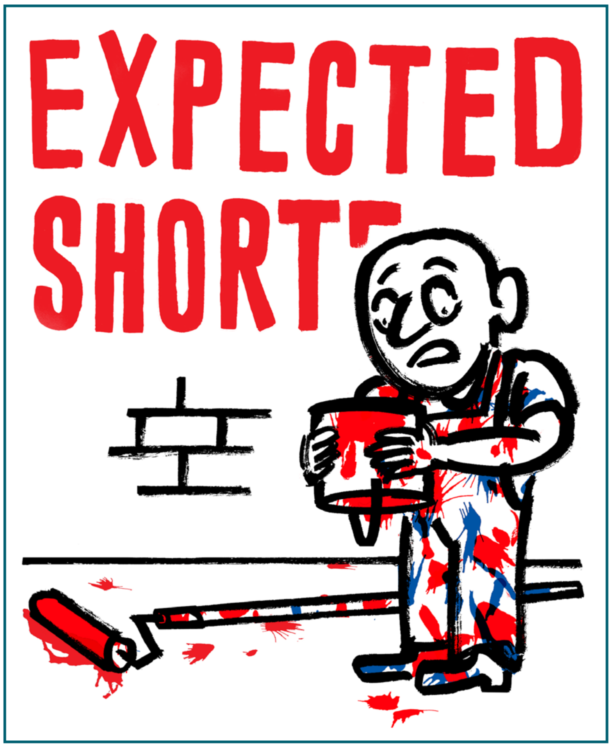


Illustration: Mark Long, nbillustration

In depth

Monthly special features: Expected shortfall

Sticking with what you know

"Quote me"

"Even a 99% VAR will be very hard to back-test unless you have very high-frequency data"
European supervisor

"My feeling is expected shortfall is not directly back-testable"
John Hull, University of Toronto

"Expected shortfall is not elicitable, which means you can't back-test it"
Gary Dunn, Morgan Stanley

"You don't have an absolute scale to measure models – the regulator should just want to make sure they are making sense over time"
Carlo Acerbi, MSCI

"It's a mathematical fact that there's no easy way of checking whether the internal model is or isn't better than some simplistic model. Without that, how do you know your model is preferable?"
Tilmann Gneiting, KIT

Eight years after the Basel Committee on Banking Supervision drew up the world's first international capital accord, the rules went through a significant revision. The *Amendment to incorporate market risks*, published in 1996, extended the framework to cover the trading book, and offered banks the choice of two approaches – a standardised framework or one based on the use of internal models. Value-at-risk was selected as the appropriate risk measure.

Regulators anticipated problems with the new framework. For example, the document warns committee members will monitor the way in which banks allocate instruments to the banking book and the trading book, and "will seek to ensure no abusive switching designed to minimise capital charges occurs".

An accompanying document laid out the terms for a back-testing regime. Banks using the internal models approach would be required to conduct quarterly back-tests of VAR at a 99% confidence interval, applying a one-day holding period. The latter condition was an attempt to assuage concerns that VAR could not be accurately tested.

Critics argued a bank's portfolio could change so much during a 10-day holding period that the resulting profit-and-loss figures would not relate to the snapshot of the portfolio taken when calculating the VAR numbers. The same charge could be laid at the door of a test based on a one-day holding period, the committee conceded, but ultimately felt a roughly accurate test was better than none at all.

In the 18 years since that document was published, much has changed. A welter of double-digit-billions trading losses in the opening phase of the financial crisis broke VAR, and the Basel Committee has since proposed it should be replaced with expected shortfall – seen as a better judge of tail risk.

But backtesting is once again a point of contention. To those who believe expected shortfall cannot be tested, Bill Coen, secretary general of the Basel Committee, says the pros of the new measure outweigh the cons. So, backtesting will continue, but it will continue to be based on VAR.

Writing in this issue, University of Toronto finance professors John Hull and Alan White describe this as "strange"; in a second article, an unnamed European regulator concedes it is "weird", but says the committee chose to stick with a process it knows.

That may be pragmatic, but it is also the kind of supervisory *non-sequitur* that gnaws away at quants and risk managers. Some researchers are now trying to come up with a robust back-testing method for expected shortfall, and it would be no surprise if regulators were quietly cheering them on.

Duncan Wood, Editor

The shortfalls of expected shortfall

Expected shortfall may be a more effective prudential measure than value-at-risk, but it is almost impossible to back-test and may be less stable than its predecessor, warn John Hull and Alan White. They also propose a simple solution for the problem created by overlapping time horizons in draft trading book rules

Need to know

- Finance professors John Hull and Alan White predict "major changes" in the calculation of market risk as a result of an ongoing overhaul of trading book capital rules.
- Among those changes, the *Fundamental review of the trading book*, currently on its second consultative paper, proposes ditching value-at-risk in favour of expected shortfall.
- This has pros and cons – during periods of stress, capital based on expected shortfall should be higher; but the capital numbers will also be less stable. Hull and White also say expected shortfall is difficult to back-test.
- Regulators acknowledge this and propose banks test VAR at two different confidence thresholds instead.
- This leads to "the strange position where the risk measure being back-tested is quite different from that used to calculate capital," the authors write.

The Basel Committee on Banking Supervision's ongoing attempts to redraw the capital rules for trading books is likely to lead to major changes in the way market risk capital is calculated.[1] After almost 20 years of using value-at-risk measures with a 10-day time horizon and a 99% confidence level, regulators have decided it is time to rethink the way capital is calculated for market risk.

There are many new approaches to calculating capital in what's known as the *Fundamental review of the trading book* (*FRTB*). We focus on two of the major changes – the switch from VAR to expected shortfall (ES), and the use of different time horizons for the shocks to market variables.[2]

ES and varying time horizons

It is proposed that VAR with a 99% confidence level be replaced by expected shortfall with a 97.5% confidence level.[3] When gains and losses are normally distributed, these two measures are almost exactly equivalent. When losses are not normally distributed, an expected shortfall with 97.5% confidence is liable to be quite a bit greater than VAR with 99% confidence. Expected shortfall in the *FRTB* is actually a stressed ES. It is to be calculated over the worst 250 days for the bank's current portfolio in recent memory.[4]

Under current regulations, VAR (when based on either current or stressed data) is calculated using a 10-day horizon. For the calculation of 10-day VAR under existing regulations, the Basel Committee allows the following formula to be used:

$$\text{10-day VAR} = \sqrt{10} \times \text{1-day VAR}$$

This means only one-day changes are considered. The formula is exactly true when daily losses – and gains – have independent normal distributions with a mean of zero and is approximately true in other situations. Under the proposed new rules, the time horizon used for a market variable will be between 10 and 250 days dependent on its liquidity. For example, a time horizon of 10 days will be used for the price of a large-cap stock while a time horizon of 120 days will be used for the credit spread of a non-investment-grade corporate.

Advantages of ES

As Artzner *et al* (1999) pointed out some time ago, ES has better theoretical properties than VAR. If two portfolios are combined, the total ES usually decreases – reflecting the benefits of diversification – and certainly never increases. By contrast, the total VAR can – and in practice occasionally does – increase. This is discussed in Hull (2006). To use the terminology of Artzner *et al*, ES is "coherent" because it has certain fundamental properties they consider such a measure should have. In particular, ES never increases as portfolios are diversified. VAR is not coherent because it does not have this particular property.

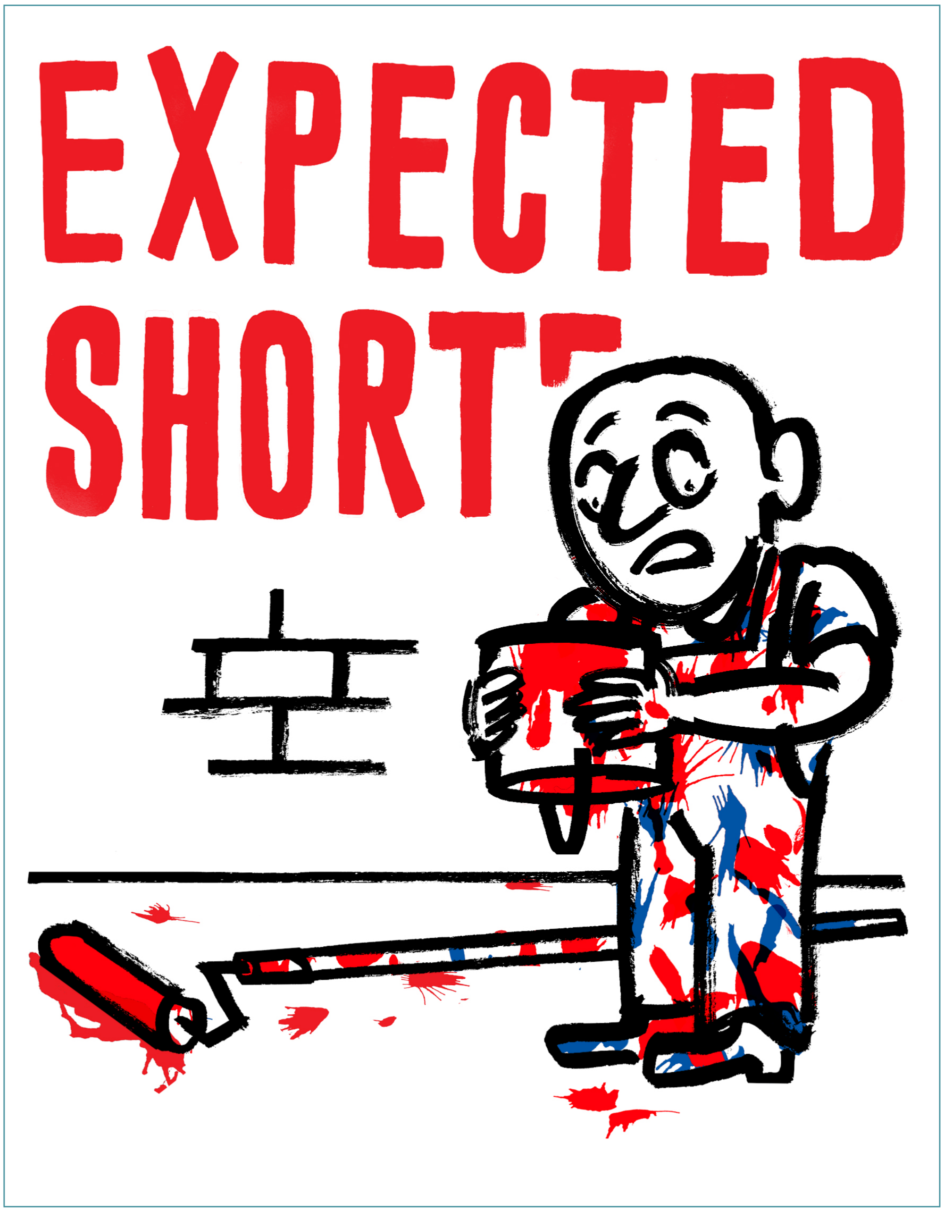


Illustration: Mark Long, nbillustration

There is a more pragmatic reason for preferring ES to VAR in risk management. It is tempting for a trader to follow a trading strategy that is nearly always profitable, but occasionally blows up.[5] This strategy should be prevented by an ES risk limit, but may be possible when a VAR risk limit is used. Many banks have used ES internally for years, even though VAR is necessary to satisfy regulatory requirements.

> It is tempting for a trader to follow a trading strategy that is nearly always profitable, but occasionally blows up.[5] This strategy should be prevented by an ES risk limit, but may be possible when a VAR risk limit is used

Back-testing and accuracy

Expected shortfall has disadvantages as well as advantages, of course. First, it is difficult to back-test (see pages 35–37). When a one-day 99% VAR model based on the most recent historical data is being back-tested, we can observe the number of exceptions that would have been encountered if the model had been used in the past, and test whether this is significantly different from what is expected. Back-testing a one-day ES model is much more challenging, because we are interested in the average size of the losses when exceptions are observed. A back-testing period of 250 days is usually used by regulators. This can be expected to give about 6 exceptions when a 97.5% confidence limit is used, which is a small sample. However, Acerbi and Szekely (2014) seem to get reasonable results when experimenting with three different tests of ES and standard distributions.

A key point is that back-testing a stressed model, whether VAR or ES, is not possible because we are interested in whether the model performs well for another stressed period, but we do not have another such period to use for testing. The use of varying time horizons in *FRTB* is an added complication in back-testing.

The Basel Committee has presumably recognised this because the review requires the back-testing of a one-day VAR model calculated in the usual way from recent historical data. We are therefore in the strange position where the risk measure being back-tested is quite different from that used to calculate capital.

Another disadvantage of ES is that estimates of the measure may not be as accurate as estimates of VAR. Yamai and Yoshiba (2002) looked at this. They found that for a certain number of observations and a certain confidence level, the accuracy of VAR and ES is about the same when the loss is normally distributed, but that VAR estimates are more accurate than ES estimates when the losses have fat tails.[6] This means capital calculated from ES may be less stable than capital calculated from VAR.

Estimating ES

The proposals in the *FRTB* recommend the use of overlapping time periods for calculations when historical simulation is used. This is markedly different from the rule mentioned above. One way a historical simulation could be carried out with overlapping time periods is as follows. In the first trial, a shock equal to the change between Day 0 and Day 10 is considered for the price of a large-cap stock, while a shock equal to the change between Day 0 and Day 120 is considered for the credit spread of a non-investment-grade corporate. Other prescribed shocks are considered for other market variables and the loss or gain in the portfolio arising from the shocks is calculated.

The second trial considers a shock equal to the change between Day 1 and Day 11 for the equity price and a shock equal to the change between Day 1 and Day 121 for the credit

Econometricians are likely to take exception to the FRTB recommendation that overlapping time periods be used. Because the changes considered when using overlapping periods are not independent, the effective sample size is much smaller than the actual sample size. As a result, although the estimate is not biased, it is very noisy

spread, and so on. The final simulation trial considers a shock equal to the change between Day 249 and Day 259 for the equity price and a shock equal to the change between Day 249 and Day 369 for the credit spread. The ES is then calculated as the average of the losses in the 2.5% tail of the distribution produced by the 250 trials.

Econometricians are likely to take exception to the *FRTB* recommendation that overlapping time periods be used. Because the changes considered when using overlapping periods are not independent, the effective sample size is much smaller than the actual sample size. As a result, although the estimate is not biased, it is very noisy. In our example, some daily credit spread changes for a non-investment-grade corporate occurring during the 250-day stressed period would be included 120 times in the 250 historical simulations. If that daily spread change was very large and positive, then it is likely that 120 samples out of the total 250 would be large and positive. The reverse would be true for a single large negative spread change.

Is there a way in which one-day changes in each market variable are used just once so that the overlapping-time-periods problem is eliminated? Suppose a non-investment-grade credit spread increases from 300 to 320 basis points in a day. What credit spread at the end of 120 days is equivalent to 320 at the end of one day? By this we mean: what percentile of the distribution of the credit spread in 120 days is the same as the percentile observed for the credit spread after one day?

One simple idea is as follows. Assume changes in the logarithm of the credit spread on successive days are independent normal distributions with zero mean and a constant standard deviation. The equivalent credit spread at the end of 120 days is:

$$300 \times \left(\frac{320}{300}\right)^{\sqrt{120}} = 608.4 \text{ basis points}$$

This estimate can be criticised in a number of ways. First, assuming the change in the logarithm of the credit spread is zero is not the same as assuming the change in the credit spread itself is zero. To correct for this, we need to know the volatility of the credit spread. Second, the volatility is not constant. If we estimate a Garch (1, 1) model, the estimate can be revised to take account of expected changes in the volatility.[7] Third, the changes in successive days may exhibit autocorrelation, with positive autocorrelation increasing the estimate, while negative autocorrelation decreases it. This can also be adjusted for.[8]

One approach to avoid different banks using different models would be for regulators, based on empirical research, to prescribe how one-day changes should be converted to the required t-day changes for the purposes of the historical simulation. A simple rule could involve setting:

$$S_{i+t} = S_i \left(\frac{S_{i+1}}{S_i}\right)^{\alpha(t)}$$

where S_i is the value of a variable on day i and $\alpha(t)$ is a parameter, possibly determined by regulators, dependent on the type of market variable being considered. More sophisticated models could also be developed.

An alternative to this is to abandon historical simulation and switch to a model-building approach in conjunction with Monte Carlo simulation. This would involve fitting a model for market variables to stressed market conditions and using it to sample changes in the variables over the prescribed number of days. In the early days of VAR, some banks used a model-building approach and some used historical simulation to calculate the measure. Eventually, historical simulation became regarded as the best approach and is now used by almost all banks. We may well go through a similar process, as banks use different approaches to implement *FRTB*. Whether historical simulation or model building emerges as the victor remains to be seen. **R**

John Hull and Alan White are professors of finance at the University of Toronto's Joseph L Rotman School of Management. The fourth edition of John Hull's book *Risk Management and Financial Institutions* will be published by Wiley in early 2015

REFERENCES

Acerbi, C. and B. Szekely (2014)
Backtesting Expected Shortfall
Risk, forthcoming

Artzner P, F Delbaen, J-M Eber and D Heath (1999)
Coherent Measures of Risk
Mathematical Finance, 9, 203-228.

Bank for International Settlements (October 2013)
Consultative Document: Fundamental Review of the Trading Book: A Revised Market Risk Framework

Hull J (2014)
Options, Futures, and Other Derivatives
Ninth edition, Upper Saddle River, NJ, Pearson

Hull J (2015)
Risk Management and Financial Institutions
Fourth edition, Hoboken, Wiley

Hull J (2006)
VAR vs. Expected Shortfall
Risk December

Yamai Y and Yoshiba T (2002)
Comparative Analysis of Expected Shortfall and Value at Risk: Their Estimation Error, Decomposition, and Optimization
Monetary and Economic Studies, January, 87–121

[1] *See Basel Committee on Banking Supervision (2013).*
[2] *See Hull (2015), chapter 17 for a more complete description of the changes proposed in the FRTB.*
[3] *ES is the expected loss conditional on the VAR level of losses being exceeded. ES is also referred to as C-VAR, conditional tail expectation, and expected tail loss.*
[4] *Currently, market risk capital is calculated as the sum of an amount based on current VAR and an amount based on stressed VAR, the latter being calculated over a similar worst-250-day period.*
[5] *A simple example of such a strategy would be the sale of deep-out-of-the money options. Strategies that work well except when there are unusual market moves, such as a 'flight to quality', can also fall into this category.*
[6] *However, the ES in FRTB has a 97.5% rather than a 99% confidence level. The lower confidence level should improve the accuracy of the ES estimate somewhat.*
[7] *See Hull (2014), chapter 23.*
[8] *See Hull (2015), chapter 12.*

VAR replacement may be too volatile, banks warn

Criticism of expected shortfall has been muted, but concerns are growing. A first-impact study showed an 80% jump in the new risk measure for one cross-asset portfolio. Cécile Sourbes **reports**

Fears are growing that expected shortfall – the mooted replacement for value-at-risk in trading book capital rules – may prove too volatile, according to Gary Dunn, a risk analytics expert at Morgan Stanley, speaking at the European regional conference of the International Swaps and Derivatives Association in London in September last year.

The new measure is a derivative of VAR that is seen as a better way of capturing tail risk. Many banks already use it for economic risk measurement purposes, so criticisms of its use in a capital context – as proposed by the Basel Committee on Banking Supervision in its Fundamental review of the trading book – have been relatively muted to date.

"I think there were so many concerns around the proposal set in the fundamental review that the industry had to prioritise its responses. But now we are getting closer to the implementation, firms themselves are looking at the volatility of this risk metric and finding that this is not necessarily the right measure for regulatory capital," said Dunn, who is the US bank's head of risk analytics for Europe, the Middle East and Africa.

A first draft of the trading book review was published in May 2012, with the intent of replacing the so-called Basel 2.5 rules – a patchwork of capital add-ons that was thrown together in a hurry following the first phase of the crisis, when losses dwarfed anything estimated by bank VAR models.

A second consultation paper appeared in October last year and the industry is now half-way through two quantitative impact studies (QISs). The results from the first, which asked banks to apply the proposed new rules to a set of hypothetical portfolios, were published on September 9. They showed an 80% jump in the risk measure for the mean of the participating banks when applied to a portfolio of mixed asset classes.

BIS headquarters, home of the Basel Committee

VAR's problem – but also its strength according to supporters – is that it only tries to measure losses up to a given confidence level and says little or nothing about a bank's exposure to more extreme events. Expected shortfall, which is the average of returns above that confidence level, has been touted as one remedy. Dunn argued these claims are now being reviewed more critically, however.

"Expected shortfall has some good features compared to VAR, which does not capture tail risk. But academics have recently been rethinking their approach and have found a few problems with this metric. First, expected shortfall is not listable, which means you can't back-test it. Second, some research suggests that since expected shortfall is an average of returns above a threshold, this makes the whole measure sensitive to extreme values. Therefore, the approach may well be volatile and is not necessarily a good metric for regulatory capital," Dunn said.

He conceded that many banks currently use expected shortfall as a risk management tool. But he warned this experience would not necessarily reveal how the measure would behave in the context of the proposed new capital framework, which calls for risk to be estimated across a series of regulator-set liquidity horizons.

Currently, banks are required to estimate potential losses on the basis of a uniform 10-day holding period – effectively, the time regulators think would be necessary to liquidate a position at something close to the prevailing market price. The trading book review would replace this with five different horizons, determined by the risk factors present in a given position. The horizons range from 10 days for a large-cap equities portfolio to 250 days for credit spreads, for instance.

"The impact of that rule is to make sure you hold some asset classes rather than others because the capital requirements are higher. But the capital requirements are higher for the less-liquid assets, which means those assets are likely to remain less liquid because the capital requirements won't make them very attractive or effective to trade," Dunn said.

Panellists were unable to provide data on the amount of capital the new regime might require.

"We are currently in the middle of the full QIS and it is clear applying the expected shortfall approach over long-term liquidity horizons will impact our model and we need to understand that. But it's too early to tell what the capital impact will be, partly because the rules still change regularly. Every couple of days, we receive a new FAQ and the rules change again, so we need to adjust our prototype," said Lars Popken, global head of market risk methodology at Deutsche Bank. **R**